DOI: 10.1214/ECP.v16-1624

Elect. Comm. in Probab. **16** (2011), 262–270

ELECTRONIC
COMMUNICATIONS
in PROBABILITY

FREEDMAN'S INEQUALITY FOR MATRIX MARTINGALES

JOEL A. TROPP[1]
Computing & Mathematical Sciences, MC 305-16, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125.
email: jtropp@acm.caltech.edu

Submitted 15 January 2011, *accepted in final form* 25 March 2011

AMS 2000 Subject classification: Primary: 60B20. Secondary: 60F10, 60G42.
Keywords: Discrete-time martingale, large deviation, probability inequality, random matrix.

Abstract

Freedman's inequality is a martingale counterpart to Bernstein's inequality. This result shows that the large-deviation behavior of a martingale is controlled by the predictable quadratic variation and a uniform upper bound for the martingale difference sequence. Oliveira has recently established a natural extension of Freedman's inequality that provides tail bounds for the maximum singular value of a matrix-valued martingale. This note describes a different proof of the matrix Freedman inequality that depends on a deep theorem of Lieb from matrix analysis. This argument delivers sharp constants in the matrix Freedman inequality, and it also yields tail bounds for other types of matrix martingales. The new techniques are adapted from recent work [Tro10b] by the present author.

1 An Introduction to Freedman's Inequality

The Freedman inequality [Fre75, Thm. (1.6)] is a martingale extension of the Bernstein inequality. This result demonstrates that a martingale exhibits normal-type concentration near its mean value on a scale determined by the predictable quadratic variation, and the upper tail has Poisson-type decay on a scale determined by a uniform bound on the difference sequence.

Oliveira [Oli10, Thm. 1.2] proves that Freedman's inequality extends, in a certain form, to the matrix setting. The purpose of this note is to demonstrate that the methods from the author's paper [Tro10b] can be used to establish a sharper version of the matrix Freedman inequality. Furthermore, this approach offers a transparent way to obtain other probability inequalities for adapted sequences.

Let us introduce some notation and background on martingales so that we can state Freedman's original result rigorously. Afterward, we continue with a statement of our main results and a presentation of the methods that we need to prove the matrix generalization.

[1]RESEARCH SUPPORTED BY ONR AWARD N00014-08-1-0883, DARPA AWARD N66001-08-1-2065, AND AFOSR AWARD FA9550-09-1-0643.

1.1 Martingales

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and let $\mathscr{F}_0 \subset \mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}$ be a filtration of the master sigma algebra. We write $\mathbb{E}_k$ for the expectation conditioned on $\mathscr{F}_k$. A *martingale* is a (real-valued) random process $\{Y_k : k = 0, 1, 2, \dots\}$ that is adapted to the filtration and that satisfies two properties:

$$\mathbb{E}_{k-1} Y_k = Y_{k-1} \quad \text{and} \quad \mathbb{E}\left|Y_k\right| < +\infty \quad \text{for } k = 1, 2, 3, \dots.$$

For simplicity, we require the initial value of a martingale to be null: $Y_0 = 0$. The *difference sequence* is the random process defined by

$$X_k = Y_k - Y_{k-1} \quad \text{for } k = 1, 2, 3, \dots.$$

Roughly, the present value of a martingale depends only on the past values, and the martingale has the status quo property: today, on average, is the same as yesterday.

1.2 Freedman's Inequality

Freedman uses a stopping-time argument to establish the following theorem for scalar martingales [Fre75, Thm. (1.6)].

Theorem 1.1 (Freedman)**.** *Consider a real-valued martingale $\{Y_k : k = 0, 1, 2, \dots\}$ with difference sequence $\{X_k : k = 1, 2, 3, \dots\}$, and assume that the difference sequence is uniformly bounded:*

$$X_k \le R \quad \textit{almost surely} \quad \textit{for } k = 1, 2, 3, \dots.$$

Define the predictable quadratic variation process of the martingale:

$$W_k := \sum\nolimits_{j=1}^{k} \mathbb{E}_{j-1}\left(X_j^2\right) \quad \textit{for } k = 1, 2, 3, \dots.$$

Then, for all $t \ge 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left\{\exists k \ge 0 : Y_k \ge t \;\; \textit{and} \;\; W_k \le \sigma^2\right\} \le \exp\left\{-\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}.$$

When the difference sequence $\{X_k\}$ consists of independent random variables, the predictable quadratic variation is no longer random. In this case, Freedman's inequality reduces to the usual Bernstein inequality [Lug09, Thm. 6].

1.3 Matrix Martingales

Matrix martingales are defined in much the same manner as scalar martingales. Consider a random process $\{\boldsymbol{Y}_k : k = 0, 1, 2, \dots\}$ whose values are matrices of finite dimension. We say that the process is a *matrix martingale* when $\boldsymbol{Y}_0 = \boldsymbol{0}$ and

$$\mathbb{E}_{k-1} \boldsymbol{Y}_k = \boldsymbol{Y}_{k-1} \quad \text{and} \quad \mathbb{E}\left\|\boldsymbol{Y}_k\right\| < +\infty \quad \text{for } k = 1, 2, 3, \dots.$$

We write $\|\cdot\|$ for the *spectral norm* of a matrix, which returns its largest singular value. As before, we define the difference sequence $\{\boldsymbol{X}_k : k = 1, 2, 3, \dots\}$ via the relation

$$\boldsymbol{X}_k = \boldsymbol{Y}_k - \boldsymbol{Y}_{k-1} \quad \text{for } k = 1, 2, 3, \dots.$$

A matrix-valued random process is a martingale if and only if we obtain a scalar martingale when we track each fixed coordinate in time.

1.4 Freedman's Inequality for Matrices

In the elegant paper [Oli10], Oliveira establishes an analog of Freedman's result in the matrix setting. He studies martingales that take self-adjoint matrix values, and he shows that the *maximum eigenvalue* of the martingale satisfies a tail bound similar to Freedman's inequality. The uniform bound R and the predictable quadratic variation $\{W_k\}$ are replaced by natural noncommutative extensions.

In this note, we establish a sharper version of Oliveira's theorem [Oli10, Thm. 1.2]. Here and elsewhere, $\lambda_{\max}$ denotes the algebraically largest eigenvalue of a self-adjoint matrix.

Theorem 1.2 (Matrix Freedman)**.** *Consider a matrix martingale $\{\boldsymbol{Y}_k : k = 0, 1, 2, \dots\}$ whose values are self-adjoint matrices with dimension d, and let $\{\boldsymbol{X}_k : k = 1, 2, 3, \dots\}$ be the difference sequence. Assume that the difference sequence is uniformly bounded in the sense that*

$$\lambda_{\max}(\boldsymbol{X}_k) \le R \quad \textit{almost surely} \quad \textit{for } k = 1, 2, 3, \dots.$$

Define the predictable quadratic variation process of the martingale:

$$\boldsymbol{W}_k := \sum\nolimits_{j=1}^{k} \mathbb{E}_{j-1}\left(\boldsymbol{X}_j^2\right) \quad \textit{for } k = 1, 2, 3, \dots.$$

Then, for all $t \ge 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left\{\exists k \ge 0 : \lambda_{\max}(\boldsymbol{Y}_k) \ge t \ \textit{ and } \ \left\|\boldsymbol{W}_k\right\| \le \sigma^2\right\} \le d \cdot \exp\left\{-\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}.$$

We prove Theorem 1.2 in Section 3 as a consequence of Theorem 2.3, a master tail bound for adapted sequences of random matrices.

Theorem 1.2 offers several concrete improvements over Oliveira's original work. His theorem [Oli10, Thm. 1.2] requires a stronger uniform bound of the form $\left\|\boldsymbol{X}_k\right\| \le R$, and the constants in his inequality are somewhat larger (but still very reasonable).

As an immediate corollary of Theorem 1.2, we obtain a result for rectangular matrices.

Corollary 1.3 (Rectangular Matrix Freedman)**.** *Consider a matrix martingale $\{\boldsymbol{Y}_k : k = 0, 1, 2, \dots\}$ whose values are matrices with dimension $d_1 \times d_2$. Let $\{\boldsymbol{X}_k : k = 1, 2, 3, \dots\}$ be the difference sequence, and assume that the difference sequence is uniformly bounded:*

$$\left\|\boldsymbol{X}_k\right\| \le R \quad \textit{almost surely} \quad \textit{for } k = 1, 2, 3, \dots.$$

Define two predictable quadratic variation processes for this martingale:

$$\boldsymbol{W}_{\mathrm{col},k} := \sum\nolimits_{j=1}^{k} \mathbb{E}_{j-1}\left(\boldsymbol{X}_j \boldsymbol{X}_j^*\right) \quad \textit{and}$$
$$\boldsymbol{W}_{\mathrm{row},k} := \sum\nolimits_{j=1}^{k} \mathbb{E}_{j-1}\left(\boldsymbol{X}_j^* \boldsymbol{X}_j\right) \quad \textit{for } k = 1, 2, 3, \dots.$$

Then, for all $t \ge 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left\{\exists k \ge 0 : \left\|\boldsymbol{Y}_k\right\| \ge t \ \textit{ and } \ \max\{\left\|\boldsymbol{W}_{\mathrm{col},k}\right\|, \left\|\boldsymbol{W}_{\mathrm{row},k}\right\|\} \le \sigma^2\right\} \le (d_1 + d_2) \cdot \exp\left\{-\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}.$$

Proof Sketch. Define a self-adjoint matrix martingale $\{\mathbf{Z}_k\}$ with dimension $d = d_1 + d_2$ via

$$\mathbf{Z}_k = \begin{bmatrix} \mathbf{0} & \mathbf{Y}_k \\ \mathbf{Y}_k^* & \mathbf{0} \end{bmatrix}.$$

Apply Theorem 1.2 to this martingale. See [Tro10b, §2.6 and §4.2] for some additional details about this type of argument. □

1.5 Tools and Techniques

In his paper [Oli10], Oliveira describes a way to generalize Freedman's stopping-time argument to the matrix setting. The main technical obstacle is to control the evolution of the moment generating function (mgf) of the matrix martingale. Oliveira accomplishes this task by refining an idea due to Ahlswede and Winter [AW02, App.]. This method, however, does not result in the sharpest bounds on the matrix mgf.

This note demonstrates that the ideas from [Tro10b] allow us to obtain sharp estimates for the mgf with minimal effort. Our main tool is a deep theorem [Lie73, Thm. 6] of Lieb.

Theorem 1.4 (Lieb, 1973)**.** *Fix a self-adjoint matrix $\mathbf{H}$. The function*

$$\mathbf{A} \longmapsto \operatorname{tr}\exp(\mathbf{H} + \log(\mathbf{A}))$$

is concave on the positive-definite cone.

See [Tro10b, §3.3] and [Tro10a] for some additional discussion of this result. We apply Theorem 1.4 through the following simple corollary [Tro10b, Cor. 3.2], which follows instantly from Jensen's inequality.

Corollary 1.5. *Let $\mathbf{H}$ be a fixed self-adjoint matrix, and let $\mathbf{X}$ be a random self-adjoint matrix. Then*

$$\mathbb{E}\operatorname{tr}\exp(\mathbf{H} + \mathbf{X}) \le \operatorname{tr}\exp(\mathbf{H} + \log(\mathbb{E}\,\mathrm{e}^{\mathbf{X}})).$$

By using Lieb's theorem instead of the Ahlswede–Winter approach, we gain a significant advantage: Our proof of Freedman's inequality can be generalized in an obvious way to obtain other probability inequalities. See the technical report [Tro11] for details.

1.6 Roadmap

Section 2 develops a master tail bound for adapted sequences of random matrices, Theorem 2.3. The proof parallels Freedman's original argument, but we require Lieb's theorem to complete one of the crucial steps. In Section 3, we specialize the main result to obtain Freedman's inequality for matrix martingales.

2 Tail Bounds via Martingale Methods

In this section, we show that Freedman's techniques extend to the matrix setting with minor (but significant) changes. The key idea is to use Corollary 1.5 to control the evolution of a matrix version of the moment generating function.

2.1 Additional Terminology

A sequence $\{X_k\}$ of random matrices is *adapted* to the filtration when each X_k is measurable with respect to $\mathscr{F}_k$. We say that a sequence $\{V_k\}$ of random matrices is *predictable* when each V_k is measurable with respect to $\mathscr{F}_{k-1}$. In particular, the sequence $\{\mathbb{E}_{k-1} X_k\}$ of conditional expectations of an adapted sequence $\{X_k\}$ is predictable. A *stopping time* is a random variable $\kappa : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ that satisfies $\{\kappa \le k\} \subset \mathscr{F}_k \quad$ for $k = 0, 1, 2, \dots, \infty$.

2.2 The Large-Deviation Supermartingale

Consider an adapted random process $\{X_k : k = 1, 2, 3, \dots\}$ and a predictable random process $\{V_k : k = 1, 2, 3, \dots\}$ whose values are self-adjoint matrices with dimension d. Suppose that the two processes are connected through a relation of the form

$$\log \mathbb{E}_{k-1} \mathrm{e}^{\theta X_k} \preccurlyeq g(\theta) \cdot V_k \quad \text{for } \theta > 0, \tag{2.1}$$

where the function $g : (0, \infty) \to [0, \infty]$. The left-hand side should be interpreted as a conditional cumulant generating function (cgf); see [Tro10b, Sec. 3.1]. It is convenient to introduce the partial sums of the original process and the partial sums of the conditional cgf bounds:

$$Y_0 := \mathbf{0} \quad \text{and} \quad Y_k := \sum\nolimits_{j=1}^{k} X_j.$$
$$W_0 := \mathbf{0} \quad \text{and} \quad W_k := \sum\nolimits_{j=1}^{k} V_j.$$

The random matrix W_k can be viewed as a measure of the total variability of the process $\{X_k\}$ up to time k. The partial sum Y_k is unlikely to be large unless W_k is also large.

To continue, we fix the function g and a positive number θ. Define a real-valued function whose two arguments are self-adjoint matrices:

$$G_\theta(Y, W) := \operatorname{tr} \exp\left(\theta Y - g(\theta) \cdot W\right).$$

We use the function G_θ to construct a real-valued random process.

$$S_k := S_k(\theta) = G_\theta(Y_k, W_k) \quad \text{for } k = 0, 1, 2, \dots. \tag{2.2}$$

This process is an evolving measure of the discrepancy between the partial sum process $\{Y_k\}$ and the cumulant sum process $\{W_k\}$. The following lemma describes the key properties of this random sequence. In particular, the average discrepancy decreases with time.

Lemma 2.1. *For each fixed $\theta > 0$, the random process $\{S_k(\theta) : k = 0, 1, 2, \dots\}$ defined in (2.2) is a positive supermartingale whose initial value $S_0 = d$.*

Proof. It is easily seen that S_k is positive because the exponential of a self-adjoint matrix is positive definite, and the trace of a positive-definite matrix is positive. We obtain the initial value from a short calculation:

$$S_0 = \operatorname{tr} \exp\left(\theta Y_0 - g(\theta) \cdot W_0\right) = \operatorname{tr} \exp(\mathbf{0}) = \operatorname{tr} \mathbf{I} = d.$$

To prove that the process is a supermartingale, we follow a short chain of inequalities.

$$\begin{aligned}
\mathbb{E}_{k-1} S_k &= \mathbb{E}_{k-1} \operatorname{tr} \exp\left(\theta \boldsymbol{Y}_{k-1} - g(\theta) \cdot \boldsymbol{W}_k + \theta \boldsymbol{X}_k\right) \\
&\le \operatorname{tr} \exp\left(\theta \boldsymbol{Y}_{k-1} - g(\theta) \cdot \boldsymbol{W}_k + \log \mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k}\right) \\
&\le \operatorname{tr} \exp\left(\theta \boldsymbol{Y}_{k-1} - g(\theta) \cdot \boldsymbol{W}_k + g(\theta) \cdot \boldsymbol{V}_k\right) \\
&= \operatorname{tr} \exp\left(\theta \boldsymbol{Y}_{k-1} - g(\theta) \cdot \boldsymbol{W}_{k-1}\right) \\
&= S_{k-1}.
\end{aligned}$$

In the second line, we invoke Corollary 1.5, conditional on $\mathscr{F}_{k-1}$. This step is legal because $\boldsymbol{Y}_{k-1}$ and $\boldsymbol{W}_k$ are both measurable with respect to $\mathscr{F}_{k-1}$. The next inequality relies on the assumption (2.1) and the fact that the trace exponential is monotone with respect to the semidefinite order [Pet94, §2.2]. The last equality is true because $\{\boldsymbol{W}_k\}$ is the sequence of partial sums of $\{\boldsymbol{V}_k\}$. □

Finally, we present a simple inequality for the function G_θ that holds when we have control on the eigenvalues of its arguments.

Lemma 2.2. *Suppose that $\lambda_{\max}(\boldsymbol{Y}) \ge t$ and that $\lambda_{\max}(\boldsymbol{W}) \le w$. For each $\theta > 0$,*

$$G_\theta(\boldsymbol{Y}, \boldsymbol{W}) \ge \mathrm{e}^{\theta t - g(\theta) \cdot w}.$$

Proof. Recall that $g(\theta) \ge 0$. The bound results from a straightforward calculation:

$$\begin{aligned}
G_\theta(\boldsymbol{Y}, \boldsymbol{W}) = \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y} - g(\theta) \cdot \boldsymbol{W}} &\ge \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y} - g(\theta) \cdot w \mathbf{I}} \\
&\ge \lambda_{\max}\left(\mathrm{e}^{\theta \boldsymbol{Y} - g(\theta) \cdot w \mathbf{I}}\right) = \mathrm{e}^{\theta \lambda_{\max}(\boldsymbol{Y}) - g(\theta) \cdot w} \ge \mathrm{e}^{\theta t - g(\theta) \cdot w}.
\end{aligned}$$

The first inequality depends on the semidefinite relation $\boldsymbol{W} \preccurlyeq w\mathbf{I}$ and the monotonicity of the trace exponential with respect to the semidefinite order [Pet94, §2.2]. The second inequality relies on the fact that the trace of a psd matrix is at least as large as its maximum eigenvalue. The third identity follows from the spectral mapping theorem and elementary properties of the maximum eigenvalue map. □

2.3 A Tail Bound for Adapted Sequences

Our key theorem for adapted sequences provides a bound on the probability that the partial sum of a matrix-valued random process is large. In the next section, we apply this result to establish a stronger version of Theorem 1.2. This result also allows us to develop other types of probability inequalities for adapted sequences of random matrices; see the technical report [Tro11] for additional details.

Theorem 2.3 (Master Tail Bound for Adapted Sequences)**.** *Consider an adapted sequence $\{\boldsymbol{X}_k\}$ and a predictable sequence $\{\boldsymbol{V}_k\}$ of self-adjoint matrices with dimension d. Assume these sequences satisfy the relations*

$$\log \mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \boldsymbol{V}_k \quad \textit{almost surely for each } \theta > 0, \tag{2.3}$$

where the function $g : (0, \infty) \to [0, \infty]$. In particular, the hypothesis (2.3) holds when

$$\mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{g(\theta) \cdot \boldsymbol{V}_k} \quad \textit{almost surely for each } \theta > 0. \tag{2.4}$$

Define the partial sum processes

$$\boldsymbol{Y}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{X}_j \quad \textit{and} \quad \boldsymbol{W}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{V}_j.$$

Then, for all $t, w \in \mathbb{R}$*,*

$$\mathbb{P}\left\{\exists k \geq 0 : \lambda_{\max}(\boldsymbol{Y}_k) \geq t \;\; \textit{and} \;\; \lambda_{\max}(\boldsymbol{W}_k) \leq w\right\} \leq d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + g(\theta) \cdot w}.$$

Proof. To begin, we note that the cgf hypothesis (2.3) holds in the presence of (2.4) because the logarithm is an operator monotone function [Bha97, Ch. V].

The overall proof strategy is the same as the stopping-time technique used by Freedman [Fre75]. Fix a positive parameter θ, which we will optimize later. Following the discussion in §2.2, we introduce the random process $S_k := G_\theta(\boldsymbol{Y}_k, \boldsymbol{W}_k)$. Lemma 2.1 states that $\{S_k\}$ is a positive supermartingale with initial value d. These simple properties of the auxiliary random process distill all the essential information from the hypotheses of the theorem.

Define a stopping time κ by finding the first time instant k when the maximum eigenvalue of the partial sum process reaches the level t even though the sum of cgf bounds has maximum eigenvalue no larger than w.

$$\kappa := \inf\{k \geq 0 : \lambda_{\max}(\boldsymbol{Y}_k) \geq t \;\; \text{and} \;\; \lambda_{\max}(\boldsymbol{W}_k) \leq w\}.$$

When the infimum is empty, the stopping time $\kappa = \infty$. Consider a system of exceptional events:

$$E_k := \{\lambda_{\max}(\boldsymbol{Y}_k) \geq t \;\; \text{and} \;\; \lambda_{\max}(\boldsymbol{W}_k) \leq w\} \quad \text{for } k = 0, 1, 2, \dots.$$

Construct the event $E := \bigcup_{k=0}^{\infty} E_k$ that one or more of these exceptional situations takes place. The intuition behind this definition is that the partial sum $\boldsymbol{Y}_k$ is typically not large unless the process $\{\boldsymbol{X}_k\}$ has varied substantially, a situation that the bound on $\boldsymbol{W}_k$ does not allow. As a result, the event E is rather unlikely.

We are prepared to estimate the probability of the exceptional event. First, note that $\kappa < \infty$ on the event E. Therefore, Lemma 2.2 provides a conditional lower bound for the process $\{S_k\}$ at the stopping time κ:

$$S_\kappa = G_\theta(\boldsymbol{Y}_\kappa, \boldsymbol{W}_\kappa) \geq \mathrm{e}^{\theta t - g(\theta) \cdot w} \quad \text{on the event } E.$$

The stopped process $\{S_{k \wedge \kappa}\}$ is also a positive supermartingale with initial value d, so

$$d \geq \liminf_{k \to \infty} \mathbb{E}[S_{k \wedge \kappa}] \geq \liminf_{k \to \infty} \mathbb{E}[S_{k \wedge \kappa} \mathbb{1}_E] \geq \mathbb{E}[\liminf_{k \to \infty} S_{k \wedge \kappa} \mathbb{1}_E] = \mathbb{E}[S_\kappa \mathbb{1}_E].$$

The indicator function decreases the expectation because the stopped process is positive. Fatou's lemma justifies the third inequality, and we have identified the limit using the fact that $\kappa < \infty$ on the event E. It follows that

$$d \geq \mathbb{E}[S_\kappa \mathbb{1}_E] \geq \mathbb{P}(E) \cdot \inf_E S_\kappa \geq \mathbb{P}(E) \cdot \mathrm{e}^{\theta t - g(\theta) \cdot w}.$$

Rearrange the relation to obtain

$$\mathbb{P}(E) \leq d \cdot \mathrm{e}^{-\theta t + g(\theta) \cdot w}.$$

Minimize the right-hand side with respect to θ to complete the main part of the argument. □

3 Proof of Freedman's Inequality

In this section, we use the general martingale deviation bound, Theorem 2.3, to prove a stronger version of Theorem 1.2.

Theorem 3.1. *Consider an adapted sequence $\{\boldsymbol{X}_k\}$ of self-adjoint matrices with dimension d that satisfy the relations*

$$\mathbb{E}_{k-1}\boldsymbol{X}_k = \mathbf{0} \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R \quad \textit{almost surely} \quad \textit{for } k = 1, 2, 3, \dots.$$

Define the partial sums

$$\boldsymbol{Y}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{X}_j \quad \textit{and} \quad \boldsymbol{W}_k := \sum\nolimits_{j=1}^{k} \mathbb{E}_{j-1}\left(\boldsymbol{X}_j^2\right) \quad \textit{for } k = 0, 1, 2, \dots.$$

Then, for all $t \ge 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left\{\exists k \ge 0 : \lambda_{\max}(\boldsymbol{Y}_k) \ge t \ \textit{ and } \ \left\|\boldsymbol{W}_k\right\| \le \sigma^2\right\} \le d \cdot \exp\left\{-\frac{\sigma^2}{R^2} \cdot h\left(\frac{Rt}{\sigma^2}\right)\right\}.$$

The function $h(u) := (1+u)\log(1+u) - u$ for $u \ge 0$.

Theorem 1.2 follows easily from this result.

Proof of Theorem 1.2 from Theorem 3.1. To derive Theorem 1.2, we note that the difference sequence $\{\boldsymbol{X}_k\}$ of a matrix martingale $\{\boldsymbol{Y}_k\}$ satisfies the conditions of Theorem 3.1 and that the martingale can be expressed in terms of the partial sums of the difference sequence. Finally, we apply the numerical inequality

$$h(u) \ge \frac{u^2/2}{1+u/3} \quad \text{for } u \ge 0,$$

which we obtain by comparing derivatives. □

3.1 Demonstration of Theorem 3.1

We conclude with the proof of Theorem 3.1. The argument depends on the following estimate for the moment generating function of a zero-mean random matrix whose eigenvalues are uniformly bounded. See [Tro10b, Lem. 6.7] for the proof.

Lemma 3.2 (Freedman mgf)**.** *Suppose that $\boldsymbol{X}$ is a random self-adjoint matrix that satisfies*

$$\mathbb{E}\boldsymbol{X} = \mathbf{0} \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{X}) \le 1.$$

Then

$$\mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{X}} \preccurlyeq \exp\left((\mathrm{e}^{\theta} - \theta - 1) \cdot \mathbb{E}(\boldsymbol{X}^2)\right) \quad \textit{for } \theta > 0.$$

The main result follows quickly from this lemma.

Proof of Theorem 3.1. We assume that $R = 1$; the general result follows by re-scaling since $\boldsymbol{Y}_k$ is 1-homogeneous and $\boldsymbol{W}_k$ is 2-homogeneous. Invoke Lemma 3.2 conditionally to see that

$$\mathbb{E}_{k-1}\,\mathrm{e}^{\theta\boldsymbol{X}_k} \preccurlyeq \exp\left(g(\theta) \cdot \mathbb{E}_{k-1}\left(\boldsymbol{X}_k^2\right)\right) \quad \text{where } g(\theta) := \mathrm{e}^{\theta} - \theta - 1.$$

Theorem 2.3 now implies that

$$\mathbb{P}\left\{\exists k \geq 0 : \lambda_{\max}(\boldsymbol{Y}_k) \geq t \ \text{ and } \ \lambda_{\max}(\boldsymbol{W}_k) \leq \sigma^2\right\} \leq d \cdot \inf_{\theta>0} \mathrm{e}^{-\theta t + g(\theta)\cdot\sigma^2}.$$

The infimum is achieved when $\theta = \log(1 + t/\sigma^2)$. Finally, note that the norm of a positive-semidefinite matrix, such as $\boldsymbol{W}_k$, equals its largest eigenvalue. □

Acknowledgments

Roberto Oliveira introduced me to Freedman's inequality and encouraged me to apply the methods from [Tro10b] to study the matrix extension of Freedman's result. I would also like to thank Yao-Liang Yu, who pointed out an inconsistency in the proof of Theorem 2.3 and who proposed the argument in Lemma 3.2. Richard Chen and Alex Gittens have helped me root out (numerous) typographic errors.

References

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, Mar. 2002. MR1889969

[Bha97] R. Bhatia. *Matrix Analysis*. Number 169 in Graduate Texts in Mathematics. Springer, Berlin, 1997. MR1477662

[Fre75] D. A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, Feb. 1975. MR0380971

[Lie73] E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.*, 11:267–288, 1973. MR0332080

[Lug09] G. Lugosi. Concentration-of-measure inequalities. Available at `http://www.econ.upf.edu/~lugosi/anu.pdf`, 2009.

[Oli10] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at `arXiv:0911.0600`, Feb. 2010.

[Pet94] D. Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory*, volume 30 of *Banach Center Publications*, pages 287–298, Warsaw, 1994. Polish Acad. Sci. MR1285615

[Tro10a] J. A. Tropp. From the joint convexity of quantum relative entropy to a concavity theorem of Lieb. Available at `arXiv:1101.1070`, Dec. 2010.

[Tro10b] J. A. Tropp. User-friendly tail bounds for sums of random matrices. Available at `arXiv:1004.4389`, Apr. 2010.

[Tro11] J. A. Tropp. User-friendly tail bounds for matrix martingales. ACM Report 2011-01, California Inst. Tech., Pasadena, CA, Jan. 2011.